

Dynamic Watermarking for General LTI Systems

Pedro Hespanhol, Matthew Porter, Ram Vasudevan, and Anil Aswani

Abstract—Detecting attacks in control systems is an important aspect of designing secure and resilient control systems. Recently, a dynamic watermarking approach was proposed for detecting malicious sensor attacks for SISO LTI systems with partial state observations and MIMO LTI systems with a full rank input matrix and full state observations; however, these previous approaches cannot be applied to general LTI systems that are MIMO and have partial state observations. This paper designs a dynamic watermarking approach for detecting malicious sensor attacks for general LTI systems, and we provide a new set of asymptotic and statistical tests. We prove these tests can detect attacks that follow a specified attack model (more general than replay attacks), and we also show that these tests simplify to existing tests when the system is SISO or has full rank input matrix and full state observations. The benefit of our approach is demonstrated with a simulation analysis of detecting sensor attacks in autonomous vehicles. Our approach can distinguish between sensor attacks and wind disturbance (through an internal model principle framework), whereas improperly designed tests cannot distinguish between sensor attacks and wind disturbance.

I. INTRODUCTION

Secure and resilient control requires the development of mechanisms to allow safe operation in the face of malicious attacks or external interferences. This is particularly challenging for cyber-physical systems (CPS) that feature interconnection between physical sensors and actuators with the communication and computation capabilities of routers, servers, etc. Such concerns are motivated by real-world instances of attacks on CPS, including: the Maroochy-Shire incident [1], the Stuxnet worm [2], and other incidents [3].

For control systems, two possible modes of attacks are either an attacker inserting faulty measurements into the output sensor signal or an attacker inserting malicious values into the actuator input for the control system. Cybersecurity techniques [4], [5], [6], [7] are an important component of designing resilient CPS. However, CPS frequently has a decentralized structure; and so approaches that detect attacks by decoupling different sensing and actuating components of the system are particularly useful for ensuring safe operation.

This paper designs a dynamic watermarking approach for detecting malicious sensor attacks for general LTI systems, and has two main contributions: First, we generalize the watermarking approach developed in [8] for SISO LTI systems

with partial state observations and MIMO LTI systems with a full rank input matrix and full state observations under an arbitrary attack, and our generalization applies to general LTI systems under a specific attack model that is more general than replay attacks [9]. Second, we show that modeling is important for designing watermarking techniques: For instance, dynamic watermarking was used to detect sensor attacks in an intelligent transportation system [10]; however, here we show that persistent disturbances such as those from wind can invalidate watermarking approaches, and we propose an approach based on the internal model principle to compensate for persistent disturbances. This second contribution motivates our generalization of dynamic watermarking to general MIMO LTI systems with partial observations, since internal model states are never directly observed.

A. Watermarking for CPS

Defense and security for CPS is classified into either detection and identification [3], [11], [12], and a number of “passive” techniques have been proposed. State estimation algorithms [13], [14] have been suggested in order to handle attacks on the physical plants within a CPS. Another related approach [15] provides a metric to characterize the resilience of a system facing stealthy attacks on the actuators.

More recently, “active defense” based on watermarking has been developed for detecting sensor attacks [8], [10], [16], [17], [18], [9], [19]. The idea is that honest (i.e., not compromised by an attacker) actuators superimpose a random signal onto the control input to ensure security in face of sensor attacks. One set of approaches [16], [17], [18], [9], [19] develops statistical hypothesis tests that detect attacks with a certain error rate, while dynamic watermarking approaches [8], [10] develop a test to ensure that only attacks which add a zero-average-power signal to the sensor measurements can remain undetected. The first set of techniques applies to general LTI systems under specific attack models, but cannot ensure the zero-average-power property for attacks; while the second set of techniques applies to specific LTI systems under general attack models. Our first contribution in this paper is to partially bridge the gap between these two techniques by developing a method that applies to general LTI systems under specific attack models and that ensures the zero-average-power property for attacks.

B. Security for Intelligent Transportation Systems

The design of intelligent transportation systems (ITS) is receiving increased attention [10], [20], [21], [22], [23], [24], [25], and one significant area for further study is the design of methods to ensure the safe and resilient operation of

This work was supported by the UC Berkeley Center for Long-Term Cybersecurity, and by Ford Motor Company.

Pedro Hespanhol and Anil Aswani are with the Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720, USA pedrohespanhol@berkeley.edu, aaswani@berkeley.edu

Matthew Porter and Ram Vasudevan are with the Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI 48109, USA matthepo@umich.edu, ramv@umich.edu

ITS. One recent work [10] considered the use of dynamic watermarking to detect sensor attacks in a network of autonomous vehicles coordinated by a supervisory controller; the watermarking approach was successfully able to detect attacks. However, large-scale deployments of ITS must be resilient in the face of persistent disturbances from environmental and human factors. Wind is an example of such a persistent disturbance. A second contribution of this work is from the perspective of modeling: We show that persistent disturbances such as those from wind can invalidate watermarking approaches, and we propose an internal model principle-based approach to handle persistent disturbances. This motivates our generalization of dynamic watermarking to general MIMO LTI systems with partial observations, since internal model states are not directly observed.

C. Outline

Section II reviews the general LTI system model (i.e., MIMO systems with partial observations) and specifies our attack model, and Sect. III provides intuition on why existing dynamic watermarking approaches cannot be used on a general LTI system. We construct a detection consistent dynamic watermarking approach for general LTI systems under our attack model in Sect. IV, and our term *detection consistent* test is used to refer to a test that ensures the zero-average-power property (described above) for attacks. Next, Sect. V describes how our asymptotic tests can be converted into statistical tests, and Sect. VI shows how our tests are special cases of those in [8] for the SISO case or the MIMO case with full rank input matrix and full state observations. We conclude with Sect. VII, which conducts simulations of an autonomous vehicle: Our tests are able to distinguish between sensor attacks and wind disturbances when including wind disturbance in the system dynamics using the internal model principal, while improperly designed tests cannot distinguish between attacks and wind.

II. LTI SYSTEM AND ATTACK MODEL

Let $[r] = \{1, \dots, r\}$, and consider a MIMO LTI system $x_{n+1} = Ax_n + Bu_n + w_n$ with partial observations $y_n = Cx_n + z_n + v_n$, where $x \in \mathbb{R}^p$, $u \in \mathbb{R}^q$, and $y, z, v \in \mathbb{R}^m$. The v_n should be interpreted as an additive measurement disturbance added by an attacker, while w_n represents zero mean i.i.d. process noise with a jointly Gaussian distribution and covariance Σ_W , and z_n represents zero mean i.i.d. measurement noise with a jointly Gaussian distribution and covariance Σ_Z . We further assume the process noise is independent of the measurement noise, that is w_n for $n \geq 0$ is independent of z_n for $n \geq 0$.

If (A, B) is stabilizable and (A, C) is detectable, then a stabilizing output-feedback controller can be designed when $v_n \equiv 0$ using an observer and the separation principle. Let K be a constant state-feedback gain matrix such that $A + BK$ is Schur stable, and let L be a constant observer gain matrix such that $A + LC$ is Schur stable. The idea of dynamic watermarking in this context will be to superimpose a private (and random) excitation signal e_n known in value to the

controller but unknown in value to the attacker. As a result, we will apply the control input $u_n = K\hat{x}_n + e_n$, where \hat{x}_n is the observer-estimated state and e_n are i.i.d. Gaussian with zero mean and constant variance Σ_E fixed by the controller.

Let $\tilde{x}^\top = [x^\top \ \hat{x}^\top]$, and define $\underline{B}^\top = [B^\top \ B^\top]$, $\underline{C} = [C \ 0]$, $\underline{D}^\top = [\mathbb{I} \ 0]$, $\underline{L}^\top = [0 \ -L^\top]$, and

$$\underline{A} = \begin{bmatrix} A & BK \\ -LC & A + BK + LC \end{bmatrix} \quad (1)$$

Then the closed-loop system with private excitation is given by $\tilde{x}_{n+1} = \underline{A}\tilde{x}_n + \underline{B}e_n + \underline{D}w_n + \underline{L}(z_n + v_n)$. If we define the observation error $\delta = \hat{x} - x$, then with the change of variables $\tilde{x}^\top = [x^\top \ \delta^\top]$ we have the dynamics $\tilde{x}_{n+1} = \underline{\underline{A}}\tilde{x}_n + \underline{\underline{B}}e_n + \underline{\underline{D}}w_n + \underline{\underline{L}}(z_n + v_n)$, where $\underline{\underline{B}}^\top = [B^\top \ 0]$, $\underline{\underline{D}}^\top = [\mathbb{I} \ -\mathbb{I}]$, $\underline{\underline{L}} = \underline{L}$, and

$$\underline{\underline{A}} = \begin{bmatrix} A + BK & BK \\ 0 & A + LC \end{bmatrix}. \quad (2)$$

Recall that $\underline{\underline{A}}$ is Schur stable whenever $A + BK$ and $A + LC$ are both Schur stable.

Since the controller is fixed, we can suppose the attacker chooses $v_n = \alpha(Cx_n + z_n) + C\xi_n + \zeta_n$ for some fixed $\alpha \in \mathbb{R}$, where $\xi_{n+1} = (A + BK)\xi_n + \omega_n$, ζ_n are i.i.d. Gaussian with zero mean and constant variance Σ_S fixed by the attacker, and ω_n are i.i.d. Gaussian with zero mean and constant variance Σ_O fixed by the attacker. The idea underlying this attack model is that the attacker allows some fraction of the true output $Cx_n + z_n$ to be measured by the controller, and at the same time also incorporates the measurement of a false state ξ_n that evolves according to the dynamics that would be expected under the controller.

III. INTUITION FOR DESIGNING A NEW TEST

To better understand how to design a new test, it is instructive to apply existing dynamic watermarking schemes and the associated tests [8] to particular LTI systems. Such an exercise provides intuition that we use to design new tests. Our main example is an LTI system with

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \text{and } C = [1 \ 0]. \quad (3)$$

Suppose the attacker chooses $v_n = -(Cx_n + z_n) + C\xi_n + \zeta_n$ with $\Sigma_S = \Sigma_Z$ and $\Sigma_O = \Sigma_W$, meaning the output measurement $y_n = C\xi_n + \zeta_n$ has no component from the actual system. This is a SISO (i.e., $m = q = 1$) system with partial state measurement, and the tests in [8] pass for this example, even though the sensor has been compromised by an attacker. The problem in this example is that the test

$$\text{as-lim}_N \frac{1}{N} \sum_{n=0}^{N-1} L(C\hat{x}_n - y_n)e_{n-1}^\top = 0 \quad (4)$$

from [8] correlates the innovations process $L(C\hat{x}_n - y_n)$ with the private excitation only one step back in time e_{n-1} ; however, it takes two time steps for the control input to enter into the output in this example. And so when designing a new test for general LTI systems, we need to take into consideration that there is generally some delay between when some private excitation is applied to when it is observed.

IV. DETECTION CONSISTENT TEST

Now let Σ_X be the positive semidefinite matrix that solves the following

$$\Sigma_X = \underline{A}\Sigma_X\underline{A}^\top + \underline{B}\Sigma_E\underline{B}^\top + \underline{D}\Sigma_W\underline{D}^\top + \underline{L}\Sigma_Z\underline{L}^\top. \quad (5)$$

Note that $\Sigma_X = \text{as-lim}_N \frac{1}{N} \sum_{n=0}^{N-1} \tilde{x}_n \tilde{x}_n^\top$. Similarly let Σ_Δ be the positive semidefinite matrix that solves the following

$$\Sigma_\Delta = (A + LC)\Sigma_\Delta(A + LC)^\top + \Sigma_W + L\Sigma_ZL^\top. \quad (6)$$

Note $\Sigma_\Delta = \text{as-lim}_N \frac{1}{N} \sum_{n=0}^{N-1} \delta_n \delta_n^\top$ and $\Sigma_\Delta = \underline{M}\Sigma_X\underline{M}^\top$, where $\underline{M} = \begin{bmatrix} 0 & \mathbb{I} \end{bmatrix}$. Recall that Σ_X and Σ_Δ exist because the above are Lyapunov equations with matrices \underline{A} , $(A + LC)$ that are Schur stable.

Lemma 1: We have that

$$\underline{A}^r \underline{B} = \begin{bmatrix} (A + BK)^r B \\ (A + BK)^r B \end{bmatrix} \quad (7)$$

for all $r \geq 0$

Proof: The result holds for $r = 0$ since $\underline{A}^0 = \mathbb{I}$ and $(A + BK)^0 = \mathbb{I}$. Now suppose the result holds for r : We prove that it holds for $r + 1$. In particular, note that

$$\underline{A}^{r+1} \underline{B} = \underline{A} \begin{bmatrix} (A + BK)^r B \\ (A + BK)^r B \end{bmatrix} = \begin{bmatrix} (A + BK)^{r+1} B \\ (A + BK)^{r+1} B \end{bmatrix}, \quad (8)$$

where the first equality holds by the inductive hypothesis, and the second equality follows by calculation of the matrix multiplication. Hence the result follows by induction. ■

Proposition 1: Let $\underline{A}(\alpha) = \underline{A} + \alpha \underline{H}$ with

$$\underline{H} = \begin{bmatrix} 0 & 0 \\ -LC & 0 \end{bmatrix}, \quad (9)$$

and define $k' = \min\{k \geq 0 \mid C(A + BK)^k B \neq 0\}$. Then we have that $\underline{A}(\alpha)^k \underline{B} = \underline{A}^k \underline{B}$ for $0 \leq k \leq k'$.

Proof: If $k' = 0$, then the result holds trivially. So assume $k' \geq 1$. We have that $\underline{A}(\alpha)^0 \underline{B} = \underline{A}^0 \underline{B} = \underline{B}$ since $\underline{A}(\alpha)^0 = \underline{A}^0 = \mathbb{I}$. Now suppose $\underline{A}(\alpha)^k \underline{B} = \underline{A}^k \underline{B}$ for $0 \leq k \leq k' - 1$. But using Lemma 1 implies that

$$\begin{aligned} \underline{A}(\alpha)^{k+1} \underline{B} &= \underline{A}^{k+1} \underline{B} + \alpha \underline{H} \begin{bmatrix} (A + BK)^k B \\ (A + BK)^k B \end{bmatrix} = \\ &= \underline{A}^{k+1} \underline{B} + \alpha \begin{bmatrix} 0 \\ -LC(A + BK)^k B \end{bmatrix} = \underline{A}^{k+1} \underline{B}, \end{aligned} \quad (10)$$

where we have used that $LC(A + BK)^k B = 0$ since $k < k'$. And so the result follows by induction. ■

Now let $k' = \min\{k \geq 0 \mid C(A + BK)^k B \neq 0\}$, and consider the following tests

$$\text{as-lim}_N \frac{1}{N} \sum_{n=0}^{N-1} (C\hat{x}_n - y_n)(C\hat{x}_n - y_n)^\top = C\Sigma_\Delta C^\top + \Sigma_Z \quad (11)$$

$$\text{as-lim}_N \frac{1}{N} \sum_{n=0}^{N-1} (C\hat{x}_n - y_n)e_{n-k'-1}^\top = 0. \quad (12)$$

Theorem 1: Suppose (A, B) is stabilizable, (A, C) is detectable, Σ_E is full rank, and $k' = \min\{k \geq 0 \mid C(A + BK)^k B \neq 0\}$ exists. If the test (11)–(12) holds, then

$$\text{as-lim}_N \frac{1}{N} \sum_{n=0}^{N-1} v_n^\top v_n = 0, \quad (13)$$

meaning that v_n asymptotically has zero power.

Proof: Observe that the dynamics for \tilde{x} are given by $\tilde{x}_{n+1} = \underline{A}(\alpha) \cdot \tilde{x}_n + \underline{B}e_n + \underline{D}w_n + \underline{L}((1 + \alpha)z_n + C\xi_n + \zeta_n)$, where $\underline{A}(\alpha) = \underline{A} + \alpha \underline{H}$ with \underline{H} given in (9). Next note that a basic calculation gives

$$\begin{aligned} \tilde{x}_n &= \underline{A}(\alpha)^k \tilde{x}_{n-k} + \sum_{k'=0}^{k-1} \underline{A}(\alpha)^{k-k'-1} (\underline{B}e_{n+k'-k} + \\ &\quad \underline{D}w_{n+k'-k} + (1 + \alpha) \cdot \underline{L}z_{n+k'-k} + \\ &\quad \underline{L}C\xi_{n+k'-k} + \underline{L}\zeta_{n+k'-k}). \end{aligned} \quad (14)$$

If we define $\underline{C} = [-C \ C]$, then $C\hat{x}_n - y_n = \underline{C}\tilde{x}_n - \alpha \cdot \underline{C}\tilde{x}_n - (1 + \alpha) \cdot z_n - C\xi_n - \zeta_n$, and so for $k \in [p]$ we have

$$\begin{aligned} \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}((C\hat{x}_n - y_n)e_{n-k}^\top) &= \\ &= (\underline{C} - \alpha \cdot \underline{C}) \cdot \underline{A}(\alpha)^{k-1} \underline{B}\Sigma_E. \end{aligned} \quad (15)$$

Note that $k' \leq p - 1$ by the Cayley-Hamilton theorem. So combining Proposition 1 with (15) implies

$$\begin{aligned} \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}((C\hat{x}_n - y_n)e_{n-k'-1}^\top) &= (\underline{C} - \alpha \cdot \underline{C}) \cdot \underline{A}^{k'} \underline{B}\Sigma_E \\ &= -\alpha \cdot \underline{C} \cdot \underline{A}^{k'} \underline{B}\Sigma_E \end{aligned} \quad (16)$$

where the second equality holds by Lemma 1 and the definition of \underline{C} . Because the test (12) holds, the quantity (16) should equal 0. But since Σ_E is full rank by assumption, Sylvester's rank inequality implies $\underline{C} \cdot \underline{A}^{k'} \underline{B}\Sigma_E \neq 0$ since

$$\underline{C} \cdot \underline{A}^{k'} \underline{B} = \begin{bmatrix} 0 \\ C(A + BK)^{k'} B \end{bmatrix} \neq 0, \quad (17)$$

where the first equality holds by Lemma 1 and the definition of \underline{C} . Thus we must have $\alpha = 0$.

Next consider the expression

$$\begin{aligned} \frac{1}{N} \sum_{n=0}^{N-1} (C\hat{x}_n - y_n)(C\hat{x}_n - y_n)^\top &= \\ \frac{1}{N} \sum_{n=0}^{N-1} (C\hat{x}_n - (1 + \alpha) \cdot (Cx_n + z_n) - C\xi_n - \zeta_n) \times \\ &\quad (C\hat{x}_n - (1 + \alpha) \cdot (Cx_n + z_n) - C\xi_n - \zeta_n)^\top. \end{aligned} \quad (18)$$

We showed above that $\alpha = 0$, and so the expectation of the above expression is

$$\begin{aligned} C\Sigma_\Delta C^\top + \Sigma_Z + \Sigma_S + \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}(C\xi_n \xi_n^\top C^\top) + \\ \frac{1}{N} \sum_{n=0}^{N-1} C(A + BK)^{N-1} x_0 (C(A + BK)^{N-1} \xi_0)^\top. \end{aligned} \quad (19)$$

Since $(A + BK)$ is Schur stable, the associated property of exponential stability implies

$$\lim_N \frac{1}{N} \sum_{n=0}^{N-1} C(A + BK)^{N-1} x_0 (C(A + BK)^{N-1} \xi_0)^\top = 0 \quad (20)$$

by combining the Cauchy-Schwartz inequality with the exponential stability. However from the test (11), the expectation must equal $C\Sigma_\Delta C^\top + \Sigma_Z$ in the limit. Since all the terms in the above expectation (19) are positive semidefinite or have zero limit, this implies

$$\Sigma_S + \text{as-lim}_N \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}(C\xi_n \xi_n^\top C^\top) = 0. \quad (21)$$

Finally, consider the expression

$$\begin{aligned} \frac{1}{N} \sum_{n=0}^{N-1} v_n v_n^\top &= \frac{1}{N} \sum_{n=0}^{N-1} ((\alpha(Cx_n + z_n) + C\xi_n + \zeta_n) \times \\ &\quad (\alpha(Cx_n + z_n) + C\xi_n + \zeta_n)^\top). \end{aligned} \quad (22)$$

Since $\alpha = 0$, the expectation of the above expression is

$$\Sigma_S + \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}(C\xi_n \xi_n^\top C^\top) + \frac{1}{N} \sum_{n=0}^{N-1} C(A+BK)^{N-1} x_0 (C(A+BK)^{N-1} \xi_0)^\top. \quad (23)$$

Combining (20)–(23) implies $\text{as-lim}_N \frac{1}{N} \sum_{n=0}^{N-1} v_n v_n^\top = 0$. However, $v_n^\top v_n$ equals the sum of the diagonal entries of $v_n v_n^\top$. Thus we have $\text{as-lim}_N \frac{1}{N} \sum_{n=0}^{N-1} v_n^\top v_n = 0$. ■

Checking for the existence of $k' = \min\{k \geq 0 \mid C(A+BK)^k B \neq 0\}$ is straightforward, though it turns out we can provide some intuitive and simple sufficient conditions.

Corollary 1: Suppose (A, B) is controllable, (A, C) is observable, and Σ_E is full rank. If the test (11)–(12) holds, then

$$\text{as-lim}_N \frac{1}{N} \sum_{n=0}^{N-1} v_n^\top v_n = 0, \quad (24)$$

meaning that v_n asymptotically has zero power.

Proof: We claim that $k' = \min\{k \geq 0 \mid C(A+BK)^k B \neq 0\} \leq p-1$ exists. Indeed, since (A, B) is controllable we have that: $(A+BK, B)$ is controllable, and the controllability matrix

$$\mathfrak{C} = [B \quad (A+BK)B \quad \dots \quad (A+BK)^{p-1}B] \quad (25)$$

has $\text{rank}(\mathfrak{C}) = p$. And so by Sylvester's rank inequality, we have $\text{rank}(C\mathfrak{C}) \geq \text{rank}(C) + \text{rank}(\mathfrak{C}) - p = \text{rank}(C)$. But (A, C) is observable, and so the observability matrix

$$\mathfrak{O} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{p-1} \end{bmatrix} = \text{diag}(C, \dots, C) \begin{bmatrix} \mathbb{I} \\ A \\ \vdots \\ A^{p-1} \end{bmatrix} \quad (26)$$

has $\text{rank}(\mathfrak{O}) = p$. Again applying Sylvester's rank inequality implies $p \text{rank}(C) \geq \text{rank}(\mathfrak{O}) = p$, or equivalently that $\text{rank}(C) \geq 1$. Combining this with the earlier inequality gives $\text{rank}(C\mathfrak{C}) \geq 1$, and so $C\mathfrak{C} \neq 0$. This means $k' \leq p-1$ exists since $C\mathfrak{C}$ is a block matrix consisting of the blocks $C(A+BK)^k B$. Thus the result follows by Theorem 1. ■

V. STATISTICAL VERSION OF TEST

For the purpose of implementation, we can also construct a statistical version of our test (11)–(12). Our approach is similar to [8] in that we construct a hypothesis test by thresholding the negative log-likelihood. Before defining the test, we make the following useful observation:

Proposition 2: Let $\psi_n^\top = [(C\hat{x}_n - y_n)^\top \quad e_{n-k'-1}^\top]$. The test (11)–(12) holds if and only if the following test holds:

$$\text{as-lim}_N \frac{1}{N} \sum_{n=0}^{N-1} \psi_n \psi_n^\top = \begin{bmatrix} C\Sigma_\Delta C^\top + \Sigma_Z & 0 \\ 0 & \Sigma_E \end{bmatrix}. \quad (27)$$

Moreover, if the test (11)–(12) holds or equivalently the test (27) holds, then we have that $\text{as-lim}_N \mathbb{E}(\psi_n) = 0$.

Proof: The equivalence between (11)–(12) and (27) follows from the definition of ψ_n and of the tests. Next suppose either (equivalent) test holds: Using the dynamics on \tilde{x} we have $\mathbb{E}(\tilde{x}_{n+1}) = \underline{A}\mathbb{E}(\tilde{x}_n) + \underline{L}\mathbb{E}(v_n)$. But we have $v_n = C(A+BK)^n \xi_0 + C \sum_{k=0}^{n-1} (A+BK)^{n-k-1} \omega_k + \zeta_n$ since $\alpha = 0$ as shown in the proof of Theorem 1, and

so $\mathbb{E}(v_n) = C(A+BK)^n \xi_0$. Since $(A+BK)$ is Schur stable, we have $\lim_n \mathbb{E}(v_n) = 0$ and hence $\lim_n \mathbb{E}(\tilde{x}_n) = \underline{A} \lim_n \mathbb{E}(\tilde{x}_n)$. This means that $\lim_n \mathbb{E}(\tilde{x}_n) = 0$ since \underline{A} is full rank (which can be seen by recalling that \underline{A} is Schur stable, so cannot have any eigenvalue of exactly one, and thus $\det(s\mathbb{I} - \underline{A}) \neq 0$ for $s = 1$). Since $C\hat{x}_n - y_n = [0 \quad C] \tilde{x}_n$, we have that $\mathbb{E}(C\hat{x}_n - y_n) = 0$. This implies $\mathbb{E}(\psi_n) = 0$ since $\mathbb{E}(e_{n-k'-1}) = 0$ by construction. ■

This result implies that asymptotically the summation $S_n = \frac{1}{\ell} \sum_{n=1}^{n+\ell} \psi_n \psi_n^\top$ with $\ell \geq m+q$ has a Wishart distribution with ℓ degrees of freedom and a scale matrix that matches (27), and we use this observation to define a statistical test. In particular, we check if the negative log-likelihood

$$\mathcal{L}(S_n) = (m+q+1-\ell) \cdot \log \det S_n + \text{trace} \left(\begin{bmatrix} (C\Sigma_\Delta C^\top + \Sigma_Z)^{-1} & 0 \\ 0 & \Sigma_E^{-1} \end{bmatrix} \times S_n \right) \quad (28)$$

corresponding to this Wishart distribution and the summation S_n is large by conducting the hypothesis test

$$\begin{cases} \text{reject,} & \text{if } \mathcal{L}(S_n) > \tau(\alpha) \\ \text{accept,} & \text{if } \mathcal{L}(S_n) \leq \tau(\alpha) \end{cases} \quad (29)$$

where $\tau(\alpha)$ is a threshold that controls the false error rate α . A rejection corresponds to the detection of an attack, while an acceptance corresponds to the lack of detection of an attack. This notation emphasizes the fact that achieving a specified false error rate α (a false error in our context corresponds to detecting an attack when there is no attack occurring) requires changing the threshold $\tau(\alpha)$.

VI. RELATIONSHIP TO EXISTING TESTS

It is interesting to compare our test (11)–(12) to those designed in [8]. More specifically, [8] designed a related sequence of tests adapted to different (and less complex) assumptions about the model dynamics. We will show that our test is closely related to (and generalizes) these previous tests developed under assumptions of less complex dynamics.

The simplest test in [8] was designed for systems with direct state measurement (i.e., $C = \mathbb{I}$), no measurement error (i.e., $z_n \equiv 0$), and full rank input matrix (i.e., $\text{rank}(B) = p$). The SISO (i.e., $m = p = q = 1$) and MIMO cases were considered separately in [8], though the SISO case is a special case of the MIMO case. If we choose $L = -A$, then we have that: $y_n = x_n + v_n$, $x_{n+1} = Ax_n + BK\hat{x}_n + Be_n + w_n$, $\hat{x}_{n+1} = Ay_n + BK\hat{x}_n + Be_n$, $\Sigma_\Delta = \Sigma_W$, and $k' = 1$ since $\text{rank}(CB) = p$. So our test (11)–(12) simplifies to

$$\text{as-lim}_N \frac{1}{N} \sum_{n=0}^{N-1} (y_{n+1} - Ay_n - BK\hat{x}_n - Be_n) \times (y_{n+1} - Ay_n - BK\hat{x}_n - Be_n)^\top = \Sigma_W \quad (30)$$

$$\text{as-lim}_N \frac{1}{N} \sum_{n=0}^{N-1} (y_{n+1} - Ay_n - BK\hat{x}_n - Be_n) \times e_n^\top = 0. \quad (31)$$

This exactly matches the test designed in [8] for LTI systems with the above described properties.

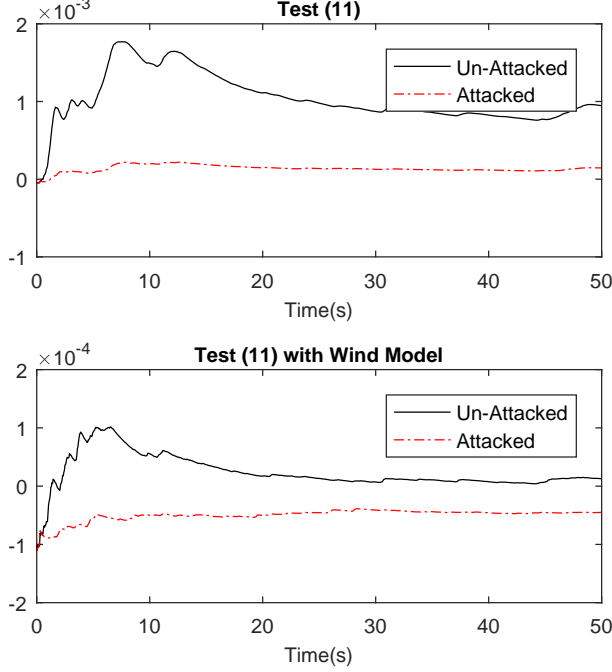


Fig. 1. Deviation of (11) in Simulation of Autonomous Vehicle

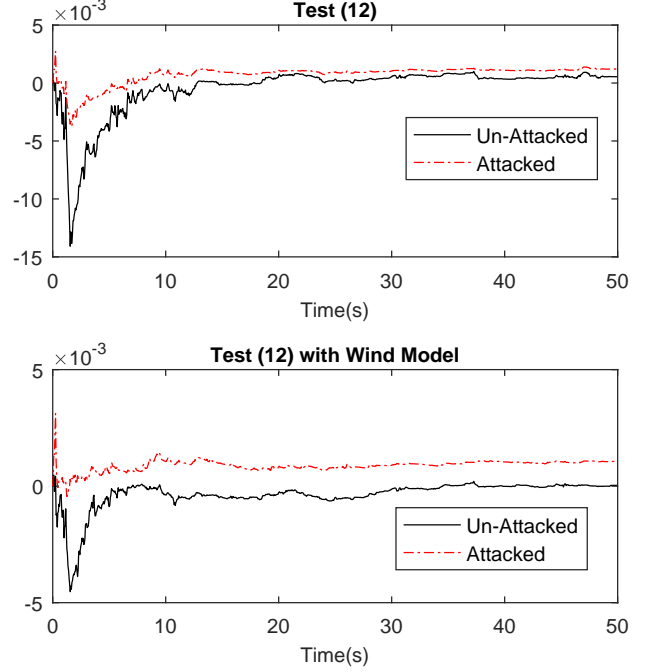


Fig. 2. Deviation of (12) in Simulation of Autonomous Vehicle

A more complex test in [8] was designed for SISO (i.e., $m = q = 1$) systems with partial state measurement. In our notation, the tests in [8] for this case simplify to

$$\text{as-lim}_N \frac{1}{N} \sum_{n=0}^{N-1} L(C\hat{x}_n - y_n)(C\hat{x}_n - y_n)^T L^T = L(C\Sigma_\Delta C^T + \Sigma_Z) L^T \quad (32)$$

$$\text{as-lim}_N \frac{1}{N} \sum_{n=0}^{N-1} L(C\hat{x}_n - y_n)e_{n-1}^T = 0. \quad (33)$$

But $k' = 1$ since B is a nonzero vector and $\text{rank}(CB) = 1$ in this case. So the test (32)–(33) from [8] essentially matches our test (11)–(12), but with the difference that the test in [8] considers quantities with $L(C\hat{x}_n - y_n)$, while our test directly considers quantities with $C\hat{x}_n - y_n$; this is a negligible difference since $C\hat{x}_n - y_n$ is a scalar in this SISO case.

VII. SIMULATIONS: AUTONOMOUS VEHICLE

A standard model [26] for error kinematics of lane keeping and speed control has $x^T = [\psi \ y \ s \ \gamma \ v]$ and $u^T = [r \ a]$, where ψ is heading error, y is lateral error, s is trajectory distance, γ is vehicle angle, v is vehicle velocity, r is steering, and a is acceleration. Linearizing about a straight trajectory and constant velocity $v_0 = 10$, and then performing exact discretization with sampling period $t_s = 0.05$ yields

$$A = \begin{bmatrix} 1 & 0 & 0 & \frac{1}{10} & 0 \\ \frac{1}{2} & 1 & 0 & \frac{1}{40} & 0 \\ 0 & 0 & 1 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad B = \begin{bmatrix} \frac{1}{400} & 0 \\ \frac{1}{2400} & 0 \\ 0 & \frac{800}{1} \\ \frac{1}{20} & 0 \\ 0 & \frac{1}{20} \end{bmatrix} \quad (34)$$

with $C = [I \ 0] \in \mathbb{R}^{3 \times 5}$. We used process and measurement noise with $\Sigma_W = 10^{-8}$ and $\Sigma_Z = 10^{-5}$, respectively.

Our simulations used the wind model: $d_{n+1} = 0.9d_n + \chi_n$, where χ_n are i.i.d. zero mean Gaussians with $\sigma_\chi^2 = 2 \times 10^{-6}$, and the wind state d entered additively into the y dynamics.

We applied our tests using a dynamic watermark with variance $\Sigma_E = \frac{1}{2}\mathbb{I}$, where K and L were chosen to stabilize the closed-loop system without an attack. We conducted four simulations: Un-attacked and attacked simulations were conducted with a test computed without wind in the system model, and un-attacked and attacked simulations were conducted with a test computed with wind in the system model. In both attack simulations, we chose an attacker with $\alpha = -0.6$, $\xi_0 = 0$, $\Sigma_O = 10^{-8}$, and $\Sigma_S = 10^{-8}$. Fig. VI shows $\|\frac{1}{N} \sum_{n=0}^{N-1} (C\hat{x}_n - y_n)(C\hat{x}_n - y_n)^T - C\Sigma_\Delta C^T - \Sigma_Z\|$, and Fig. VI shows $\|\frac{1}{N} \sum_{n=0}^{N-1} (C\hat{x}_n - y_n)e_{n-k'-1}^T\|$. If the test is detection consistent, then these values go to zero. The plots show dynamic watermarking cannot detect the presence or absence of an attack when wind affects the system dynamics but is not included in the test, while our test sdetect the presence or absence of an attack when a model of wind is included in the test. Fig. VII shows the results of applying our statistical test (28), and the same behavior is seen.

VIII. CONCLUSION

This paper constructed a dynamic watermarking approach for detecting malicious sensor attacks for general LTI systems, and the two main contributions were: to extend dynamic watermarking to general LTI systems under a specific attack model that is more general than replay attacks, and to show that modeling is important for designing watermarking techniques by demonstrating how persistent disturbances can negatively affect the accuracy of dynamic watermarking. Our

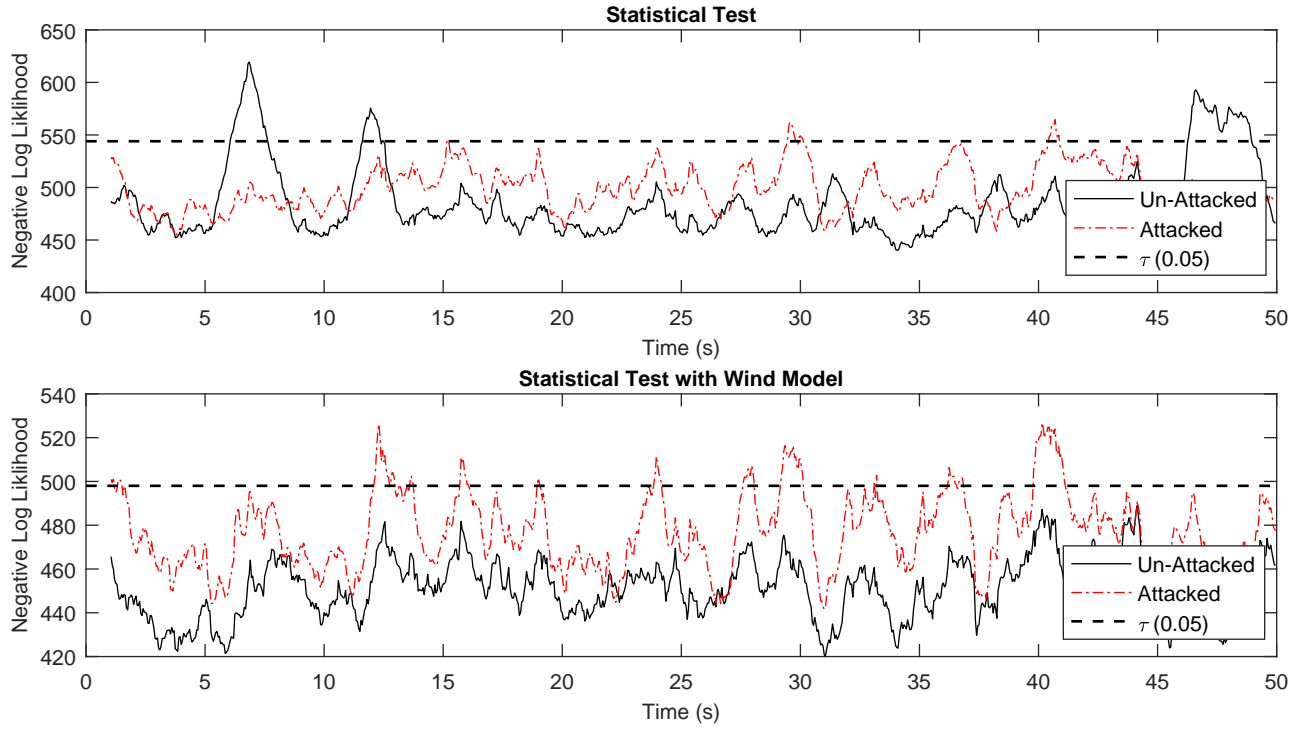


Fig. 3. Value of (28) for Simulation of Autonomous Vehicle, with a Negative Log-Likelihood Threshold for $\alpha = 0.05$ False Detection Error Rate

approach to resolve this issue was to incorporate a model of the persistent disturbance via the internal model principle. Future work includes generalizing the attack models that can be detected by our approach. That probably requires a slight modification of our test: In particular, a test applicable to an arbitrary attack likely requires that (12) hold for all $k' \in [p]$.

REFERENCES

- [1] M. Abrams and J. Weiss, "Malicious control system cyber security attack case study—Maroochy water services, australia," *MITRE*, 2008.
- [2] R. Langner, "Stuxnet: Dissecting a cyberwarfare weapon," *IEEE Security & Privacy*, vol. 9, no. 3, pp. 49–51, 2011.
- [3] A. A. Cárdenas, S. Amin, and S. Sastry, "Research challenges for the security of control systems," in *HotSec*, 2008.
- [4] B. Parno, M. Luk, E. Gaustad, and A. Perrig, "Secure sensor network routing: A clean-slate approach," in *ACM CoNEXT*, 2006.
- [5] V. Kumar, J. Srivastava, and A. Lazarevic, *Managing cyber threats: issues, approaches, and challenges*. Springer, 2006, vol. 5.
- [6] W. Wang and Z. Lu, "Cyber security in the smart grid: Survey and challenges," *Computer Networks*, vol. 57, no. 5, pp. 1344–1371, 2013.
- [7] K.-D. Kim and P. R. Kumar, "Cyber-physical systems: A perspective at the centennial," *Proc. of IEEE*, vol. 100, pp. 1287–1308, 2012.
- [8] B. Sathidhanandan and P. Kumar, "Dynamic watermarking: Active defense of networked cyber-physical systems," *Proc. of IEEE*, 2016.
- [9] S. Weerakkody, Y. Mo, and B. Sinopoli, "Detecting integrity attacks on control systems using robust physical watermarking," in *Proc. of IEEE CDC*, 2014, pp. 3757–3764.
- [10] W.-H. Ko, B. Sathidhanandan, and P. Kumar, "Theory and implementation of dynamic watermarking for cybersecurity of advanced transportation systems," in *Proc. of IEEE CNS*, 2016, pp. 416–420.
- [11] A. A. Cardenas, S. Amin, and S. Sastry, "Secure control: Towards survivable cyber-physical systems," in *ICDCS*, 2008, pp. 495–500.
- [12] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, 2013.
- [13] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1454–1467, 2014.
- [14] —, "Secure state-estimation for dynamical systems under active adversaries," in *Allerton Conference*. IEEE, 2011, pp. 337–344.
- [15] C.-Z. Bai, F. Pasqualetti, and V. Gupta, "Security in stochastic control systems: Fundamental limitations and performance bounds," in *American Control Conference (ACC)*, 2015. IEEE, 2015, pp. 195–200.
- [16] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Allerton Conference*. IEEE, 2009, pp. 911–918.
- [17] Y. Mo, E. Garone, A. Casavola, and B. Sinopoli, "False data injection attacks against state estimation in wireless sensor networks," in *Proc. of IEEE CDC*, 2010, pp. 5967–5972.
- [18] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on scada systems," *IEEE CST*, vol. 22, no. 4, pp. 1396–1407, 2014.
- [19] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs," *IEEE Control Systems*, vol. 35, no. 1, pp. 93–109, 2015.
- [20] H. Gonzalez and E. Polak, "On the perpetual collision-free rhc of fleets of vehicles," *Journal of optimization theory and applications*, vol. 145, no. 1, pp. 76–92, 2010.
- [21] A. Aswani and C. Tomlin, "Game-theoretic routing of gps-assisted vehicles for energy efficiency," in *ACC*, 2011, pp. 3375–3380.
- [22] W. Zhang, M. Kamgarpour, D. Sun, and C. J. Tomlin, "A hierarchical flight planning framework for air traffic management," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 179–194, 2012.
- [23] R. Vasudevan, V. Shia, Y. Gao, R. Cervera-Navarro, R. Bajcsy, and F. Borrelli, "Safe semi-autonomous control with enhanced driver modeling," in *ACC*, 2012, pp. 2896–2903.
- [24] S. Mohan and R. Vasudevan, "Convex computation of the reachable set for hybrid systems with parametric uncertainty," in *Proc. of ACC*, 2016, pp. 5141–5147.
- [25] G. Como, E. Lovisari, and K. Savla, "Convexity and robustness of dynamic network traffic assignment for control of freeway networks," *IFAC-PapersOnLine*, vol. 49, no. 3, pp. 335–340, 2016.
- [26] V. Turri, A. Carvalho, H. Tseng, K. Johansson, and F. Borrelli, "Linear model predictive control for lane keeping and obstacle avoidance on low curvature roads," in *Proc. of IEEE ITSC*, 2013, pp. 378–383.